MODELING OF STRATUM VARIANCE FOR USE IN SAMPLE ALLOCATION IN AGRICULTURAL AREA FRAME SURVEYS

Charles R. Perry, USDA-NASS
Raj S. Chhikara[1,2], University of Houston-Clear Lake
Floyd M. Spears[1,3], Harding University
Charles Perry, USDA-NASS, 3251 Old Lee Highway, Room 305, Fairfax, VA 22030

KEY WORDS: Composite Variance Estimation, Model Fit, Multivariate Allocation, Stratified Sample Design, Variance Function

Abstract

NASS updates its area sampling frame for its agricultural surveys on a regular basis, but its update often has a lag period, sometimes of many years. Since the use of agricultural land in an area may change between two consecutive update periods, the previous sample design stratum variance estimates may not reflect the actual variance for one or more strata. This has led NASS to consider development of an alternative method of estimating stratum variances to be used in sample allocation. A variance function originally proposed by Mahalanobis (1940) for crop acreage and by Smith (1936) for crop yield can be a basis for the development of a stratum variance model. In this study, the stratum variance is modeled for each of the eight agriculture items used in the multivariate sample allocation by NASS for its agricultural surveys.

Survey data from June 2002 are used in the stratum variance model development. Data are grouped into major land use strata and separate model-fits are made for the groups. All model-fits were performed using a weighted linear regression method and the weights were developed using an empirical approach. Sample allocation was made using both the survey variance estimates and the model predicted variances, and the two allocations were compared with the original sample allocation used by NASS for its June 2002 agricultural survey.

[1] These authors' work was supported in part under a cooperative research program of the United States Department of Agriculture at the University of Houston-Clear Lake.

[2] This author is with the University of Houston-Clear Lake, 2700 Bay Area Blvd, Houston, TX 77058.

[3] This author is with Harding University, Box 10764, Searcy, AR 72149.

1 Introduction

The National Agricultural Statistics Service (NASS) of the U.S. Department of Agriculture (USDA) uses a multi-frame sampling methodology to make estimates of agricultural items. The list frame sampling is used for making the primary estimates of all items and the area frame sampling is utilized either for the list incompleteness or the primary estimates of crop acreage. A five year rotation sample design is the basis for area frame sampling and a sub-stratification process is employed in the selection of sample units in a stratum. The measurement unit size is nominally one square mile in area in moderate to high intensity agriculture strata, and it ranges from nominally one square mile to eight square miles in area in the low intensity agricultural strata. PPS sampling is used in sparsely agricultural areas. The unit size has area 0.10 or 0.25 square miles in the urban or agriculture urban strata.

A stratified design is employed in the NASS agricultural area frame surveys. Land use stratification is carried out primarily based on recent cultivated land, pastures, forest, and other land features. This is done by delineating recognizable patterns from photo-interpretation of aerial photographs and landsat imagery. The land use stratification is updated at a regular time interval and involves an extensive use of mapping technology. A multivariate procedure is the basis of NASS area frame sample allocation. It requires input values for eight agricultural items, of which six are major crops and two are non-crop items. The six crops are corn, cotton, soybeans, durum wheat, spring wheat and winter wheat. The two non-crop items are the number of farms and the number of cattle that are not covered by the list frames, designated as the not-on-list (NOL) cattle. One potential addition is the number of equines. We will consider the case of NOL equines only for the purpose of modeling its variance or standard deviation for various groups of strata.

Determination of sample size and its allocation to different strata requires input of stratum variances for each of the eight agricultural items. NASS utilizes stratum variances determined from 3-5 years' historical average of variances computed from survey data. However, sample survey variances can be unstable when a stratum has a small number of samples or when the percent crops or agriculture items are small. Moreover, there is a lack of independence among historical estimates due to the use of rotation design sampling. So the major objective of this study is to obtain more stable estimates of stratum variances for the eight items for their use in the multivariate sample allocation.

In the present study, six crops and three non-crop items are considered in modeling of stratum variances. Upon examination of the survey estimates for the non-crop items in 2001, 2002, and 2003, it appeared that the stratum standard deviation is linearly dependent upon the item value in the stratum, and thus it may be modeled as function of the item value. However, in the case of crop acreage, the stratum variance itself could be modeled as function of the item value or its proportion in the stratum or a function of the proportion. The variance function initially used by Mahalonobis (1940) was found to be appropriate for modeling of stratum variance for each of the crops. Perry (1992) applied it successfully to model stratum variance for sample allocation of segments for major crops in Pakistan.

Empirical model-fits for stratum variance (for each crop) and stratum standard deviation (for each noncrop item) were obtained using the 2001, 2002, and 2003 NASS survey estimates. These are detailed later only for year 2002 in Section 3. The adequacy of the model-fits is examined based on observed significance levels and the weighted residual plots discussed in Section 4.

2 Data Use in Modeling

The 2001, 2002, and 2003 survey data were used in the modeling data of stratum estimates and variance estimates for the agriculture items of six crops and three non-crops as listed in Section 1. Both the estimate and the variance estimate for an item are computed for each substratum within each land use stratum. The substratum quantities are aggregated to obtain the stratum level estimates.

NASS land use strata are grouped together by considering similarity in their major land use. These strata are numbered in teens, twenties, thirties, forties, fifties, and higher depending upon the agricultural intensity or the location. Table 1 lists the characteristics of the five groups of land use strata considered for the area level at which the model-fits are made.

Table 1: Stratum Groups

Stratum Group	Land-use Strata	Description
1	10 - 19	Intense Cultivation
2	20 - 29	Moderate Cultivation
3	30 - 39	Urban or Ag Urban
4	40 - 49	Low Cultivation
5	≥ 50	Non-Agricultural

Survey data were standardized on a per acre basis. This was to eliminate the effect of segment size which varied from one group of land use strata to anther group. This amounted to using the proportion of an ag item per acre in a stratum and the corresponding survey variance estimate. Furthermore, the direct survey estimates for each item was replaced by those determined using the Agricultural Statistics Board (ASB) estimates. Since there are no ASB estimates at the stratum level, state estimates were proportioned to strata within the state. The proportioning was done based on the cultivated land in each stratum relative to the state cultivated land in the case of crop items. Table 2 lists the agricultural items used for proportioning of ASB estimates for the eight items considered here.

Table 2: Proportioning Factors for Agricultural Items

Item of Interest	Proportioning Factor
Each crop	Cultivated land
Number of farms	Land in farms
NOL cattle	Crop land
Equines	Number of farms

For land use stratum h, let x_h denote the ASB proportioned estimate of an item and s_h^2 the variance estimate computed from the survey data. Then the estimated quantities on a per acre basis are as follows:

The crop proportion in stratum $h$is

$$p_k = \frac{x_h}{(\text{Total Acreage})_h}$$

where

$$x_h = \frac{(\text{Cultivated Land})_h}{(\text{Cultivated Land})_{\text{state}}} \cdot (\text{ASB Estimate})_{\text{state}}$$

and the corresponding variance is

$$s_h^2 = \frac{(\text{Survey Variance Estimate})}{N_h^2(\text{Segment Size})^2}$$

where $N_h$is the total number of area segments in the stratum.

3 Modeling of Stratum Variance

3.1 Case of Crop Acreage

If the measurement unit is same as the sampling unit and it either is completely covered by the crop of interest or it is not, then sampling of a unit randomly in a stratum amounts to a Bernoulli trial. If p is the proportion of units having the crop of interest, then the stratum variance at the unit level is given by

$$\sigma^2 = p(1-p)$$

Since the measurements are at the aggregate level, one can postulate the stratum variance to be

$$\sigma^2 = \beta \cdot y^\alpha$$

where $y = p(1-p)$.

As explained in Mahalanobis (1944), the between area segment variance for a crop acreage can be expected to be a power function of the variance under a binomial model. The power would depend upon the intra-class correlation of the measurement units devoted to the crop in area segment and may be empirically determined using the survey estimates. This formulation was the basis of an empirical modeling of stratum variance by Mahalanobis (1940). The following model results for the stratum variance.

$$\sigma^2 = \beta y^\alpha + \epsilon$$

where y is as defined above and ϵ is the random error component. Initially a non-linear model-fit was made using NASS data. However, in almost all model-fits, the estimated value of α did not differ significantly from 1, a linear model-fit was done, assuming $\alpha = 1$.

3.2 Case of Non-Crop Items

A counting process seems applicable for the occurrence of the number of farms, cattle, or equines in an area segment. When the survey estimates of these items were examined, it was found that a fewer number of them occur much more frequently than do a higher number of them in a segment. The frequency histogram plots in Figure 1 for farms, cattle and equines suggested that an exponential distribution can be assumed for the underlying probability distribution for each item. Accordingly, the stratum standard deviation is equal to the stratum mean value, $\sigma = \mu$. Thus the following model is considered for the stratum standard deviation:

$$\sigma = \beta y + \epsilon$$

where y is the item estimate in a stratum.

3.3 Weighted Model Fitting

Scatter plots were made for stratum variance (s_h^2) in the case of crops, and standard deviation (s_h) in the case of non-crops, vs. covariate (y_h) for each items. In almost all cases, s_h^2 (or s_h) on the average was increased as $y_h$increased. However, the ranges in s_h^2 (as well as in y_h) differ considerably across the five groups of strata and thus separate model-fits were considered for various group strata.

The weighting in model fitting was investigated to account for the error variance heterogeneity, reliability of survey variance estimates and outliers in the data. The scatter plot of paired data (y_h,s_h) was made showing each point by a bubble with size proportional to sample size used in computing the paired values. This allowed us to examine the points in scatter plot according to their relative precision and hence, their importance. This led us to assign a weight to each data point based on the associated sample size in developing a model-fit.

Since the s_h^2 are computed from sample survey data, these are not equally reliably estimated. Because a larger sample size leads to higher precision for s_h, its weight is considered proportional to n_h.

Scatter plots of s_h vs. $x_h$showed that on the average the s_h value increases as $x_h$increases, and so does the spread in the s_h values. As defined previously, $x_h$denotes the stratum estimate for an item. This implies the variance of s_h increases as a function of x_h. This requires carrying out a weighted linear model-fit for s_h where the model residuals are weighted by an inverse power of x_h. We used survey data to estimate the variance of s_h based on such considerations. Detailed analysis is omitted here and

will be described in a forthcoming NASS Research Report.

In order to deal with potential outliers in data, Tukey's biweight procedure was used to assign lower weights to extreme observations. See Fox (2002) for its full description.

4 Model Fits

For each group stratum, an estimate of β, denoted as b, was determined by minimizing the following quantities:

$$\sum_{h=1}^{H}(s_h^2 - \beta y_h)^2 w_h$$

in the case of crop acreage, and

$$\sum_{h=1}^{H}(s_h - \beta y_h)^2 w_h$$

in the case of a non-crop item.

Here H represents the number of land use strata in a group stratum and $y_h$represents the covariate value as defined earlier.

Table 3 lists the estimates of β and other statistics for model fits for the crop and non-crop items.

A weighted residual plot was made and examined for any lack of fit or anomaly in each model-fit. The model-fits were viewed to be reasonable to good for group strata 1 and 2, but not so good for group strata 3 and 4. Since group strata 1 and 2 account for a substantially large amount of value for an item, the model-fits were judged to be useful for predicting stratum variances.

5 Sample Allocation

NASS's multivariate allocation procedure was utilized with the stratum variances obtained for the eight agricultural items of six crops and two non-crops. Equines were not included since there were no input values available for the constraint needed in the allocation procedure. For comparison purposes, we considered the use of three different stratum variances to determine the sample size and its allocation to land use strata:

1. S_D^2 = Direct Survey Variance Estimate
2. S_M^2 = Model-Fit Predicted Variance
3. $S^2 = \alpha S_D^2 + (1-\alpha)S_M^2$, a composite variance estimate with α defined as:

$$\alpha = \begin{cases} 1 & \text{if} df \geq 30 \\ \frac{df}{30} & \text{if} df < 30 \end{cases}$$

Here df = n-k, the degrees of freedom (df) associated with s^2, where n=number of sample segments, and k=number of substrata in a land use stratum.

Since s_D^2 is a pooled variance for the stratum, its reliability depends upon the degrees of freedom for the within sum of squares computed across substrata in the stratum. So the reasoning behind this definition of α is that the stratum variance estimate is expected to be reliable when the degrees of freedom is equal to 30 or more. Otherwise, it needs to be weighted down depending upon its reliability relative to the case of df=30.

Table 4 lists the sample allocations obtained in each case. For comparison, it also lists the original allocation that NASS used in its June 2002 agricultural survey.

Figure 2 shows a plot of stratum sample allocations obtained using the composite variance estimates versus their original sample allocation across all group strata. The new sample allocation is quite in agreement with the original allocation except in a few cases. Almost all discrepancies between the two allocations occur for land use strata in group stratum 3 and 4. The new allocation tended to assign a few more samples to strata in group stratum 4. In a few cases, land use strata from group stratum 1 were assigned a fewer number of samples. For example, these discrepancies can be seen in Figure 2 for the case of "Allocation < 50".

6 Conclusions

For the three non crop items, the weighted linear model-fit for the stratum standard deviation as a function of the stratum item value has very small observed significance levels (p-values). This and the weighted residual plots indicate the model-fits are reasonably good and can be effectively used in predicting the stratum standard deviation for each of these items.

Similarly for the crop acreage, the weighted model-fits are found to be adequate in the prediction of stratum variances for crop acreage for all crops. However, the model-fits for group stratum 3 are not significant for cotton and spring wheat and hence, not reliable in predicting stratum variance.

There are substantially less allocated samples in Group Stratum 1 and 2, and more samples in Group Stratum 4 when model predicted variances are used. When the composite stratum variances are used, the

Table 3: Model Fits for Crop and NonCrop Items

Crop Items: Stratum Variance Model-Fits

Item	Stratum Group	Number of Land Use Strata	b	SE(b)	P-Value
Corn	1	57	0.153862	0.012954	6.24E-17
Corn	2	50	0.07781	0.004583	3.64E-22
Corn	3	100	0.217178	0.030183	1.21E-10
Corn	4	46	0.098786	0.013787	5.8E-09
Cotton	1	19	0.218772	0.024609	4.1E-07
Cotton	2	24	0.097984	0.024605	0.000588
Cotton	3	35	0.000444	0.002014	0.826877
Cotton	4	18	0.010502	0.003933	0.016152
Durum Wheat	1	36	0.018871	0.007359	0.014789
Soybeans	1	35	0.171235	0.014735	8.15E-13
Soybeans	2	35	0.118513	0.012649	6.02E-11
Soybeans	3	69	0.129221	0.040361	0.002079
Soybeans	4	35	0.083169	0.013164	3.34E-07
Spring Wheat	1	14	0.159423	0.017146	2.64E-07
Spring Wheat	2	16	0.146406	0.01224	2.65E-09
Spring Wheat	3	29	0.046378	0.045021	0.311769
Spring Wheat	4	8	0.013672	0.000415	6.16E-09
Winter Wheat	1	48	0.23557	0.021038	1.24E-12
Winter Wheat	2	52	0.107688	0.014348	8.67E-10
Winter Wheat	3	93	0.264018	0.06328	6.83E-05
Winter Wheat	4	41	0.017397	0.001882	1.16E-10

Non-Crop Items: Stratum Standard Deviation Model-Fits

Item	Stratum Group	Number of Land Use Strata	b	SE(b)	P-Value
Cattle	1	76	0.049467	0.008186	1.76E-07
Cattle	2	54	0.162095	0.024427	0.000139
Cattle	3	100	0.368933	0.031187	0.00289
Cattle	4	47	0.351946	0.048941	1.6E-206
Equines	1	78	1.526822	0.124748	9.16E-14
Equines	2	54	1.441874	0.113686	3.72E-10
Equines	3	100	1.67626	0.231321	1.1E-14
Equines	4	47	1.17125	0.195901	3.38E-07
Farms	1	78	0.605842	0.046503	8.96E-15
Farms	2	56	0.808442	0.043677	3.97E-20
Farms	3	100	1.059847	0.069307	7.27E-26
Farms	4	47	1.260453	0.148269	2.7E-10
Farms	5	15	2.973199	9.81E-17	4E-224

Table 4: Sample Allocation Comparisons

Stratum Group	Original Allocation	Allocation Obtained Using		
		2002 Survey Data	Model Predicted	Composite Variance
1	6389	6268	5627	6458
2	2454	2400	2174	2490
3	284	285	270	271
4	955	927	1345	1035
5	96	96	96	96
Total	10,178	9976	9512	10,350

sample allocation for 2002 is relatively unchanged compared to the original allocation except for Group Stratum 4 which gets assigned more number of samples to its land use strata.

References

[1] Allen, Don (1999), "1997 Census Not on Mail List Survey". U.S. Department of Agriculture, National Agricultural Statistics Service Report (unpublished).

[2] Bethel, James (1989), "Sample Allocation in Multivariate Surveys," *Survey Methodology*, **15:1**, 47-57.

[3] Chhikara, Raj S. and Perry, Charles R. (1986), "Estimation of Stratum Variance in Planning of Crop Acreage Surveys"' *Journal of Statistical Planning and Inference*, **15**, 97-114.

[4] Chhikara, Raj S., Spears, Floyd M. and Perry, Charles R. (2002), "Sample Allocation for Estimation of the Number of "Not on Mail List" (NML) Farms for the 2002 Census of Agriculture," USDA-NASS, RDD Research Report Number RDD-02-01, January, 2002.

[5] Fox, John (2002), *Robust Regression,* Appendix to "An R and S-PLUS Companion to Applied Regression." January, 2002.

[6] Fuller, Wayne A. (1987), *Measurement Error Models,* Wiley, New York.

[7] Mahalonobis, P.C. (1940), "A Sample Survey of the Acreage Under Jute in Bengal," *Sankhyā*, **4**, 511-530.

[8] Mahalonobis, P.C. (1944), "On Large Scale Sample Surveys," *Philosophical Transactions of Royal Society, Landon, Series B*, **231**, 329-451.

[9] National Agricultural Statistics Service (1989), *Area Frame Design for Agricultural Surveys*, Washington, D.C.: NASS, U.S. Department of Agriculture.

[10] Perry, Charles R. (1992), "The Province Level Sample Size and Allocation for the Punjab, Sindh and Northwest Frontier Provinces," IPO Trip Report, NASS, USDA. Washington, D.C. December, 1992.

[11] Smith, H.F. (1936), "An Empirical Law Describing Heterogeneity in the Yield of Agricultural Crops," *Journal of Agricultural Science*: **28**.

Figure 1: Occurrences per Segment for Farms, NOL Cattle and Equines in 2002

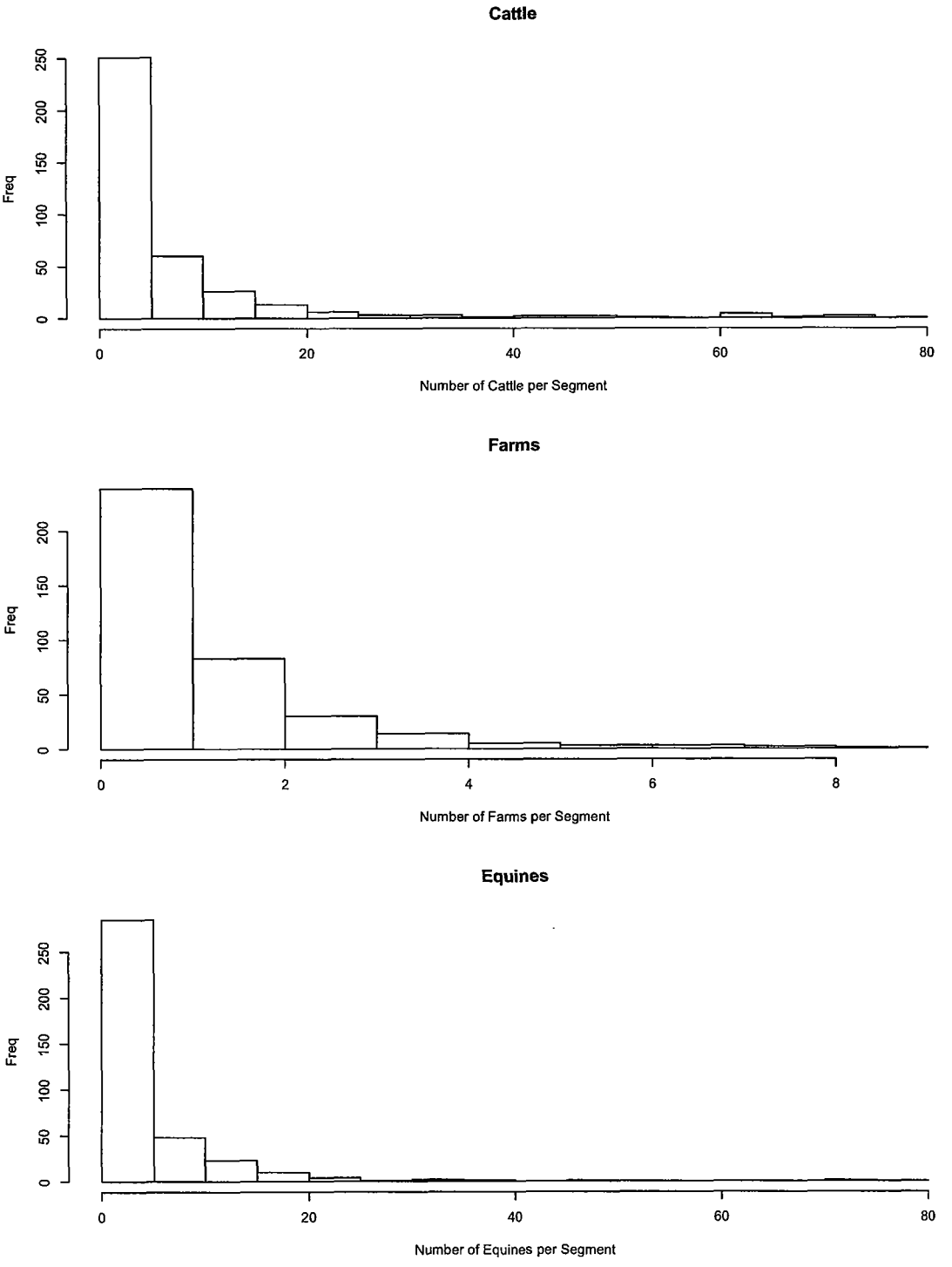

Figure 2: NASS Original Allocation versus Composite Allocation for 2002